Genomic Data Access Through BLAST

https://blast.ncbi.nlm.nih.gov

Accessing genomic sequence data through BLAST, on the web or using standalone tools

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Introduction

Advances in next generation sequencing technology (NGS) have led to the availability of genomic sequence data for an increasing large number of organisms. BLAST searching against these datasets, particularly the annotated assemblies based on raw sequence reads, can provide significant insight into the biology of these biomedically, agriculturally and ecologically important species. However, assembly and annotation from raw sequence reads is a complex process. The availability of best sequence data vary from organism to organism and will require different access strategies. In this document, we will go over the organization of genomic sequence data available from NCBI, and ways to locate the best genomic dataset for the organism of interest for sequence alignment purposes, through both the BLAST homepage at blast.ncbi.nlm.nih.gov/, other web pages/records at NCBI, as well as through standalone tools provided by NCBI.

Workflow and Organization for Nucleotide Sequence Data

From an NGS-centric point of view, we can organize available nucleotide sequence data based on their volume, degree of assembly, and information density (quality of annotation) as a pyramid structure (Figure 1, right). The figure separates nucleotide sequences into genomic and transcripts entries (left and right), and sorts them by their level of assembly (bottom to top). We should use the records at the top of the pyramid with the highest level of assembly and annotation. It is better to access the data through the organizational databases, e.g., Assembly or BioProject (left half), since they organize and connect related nucleotide records, such as individual chromosomes for a specific organism, into a biologically-relevant collection unit.

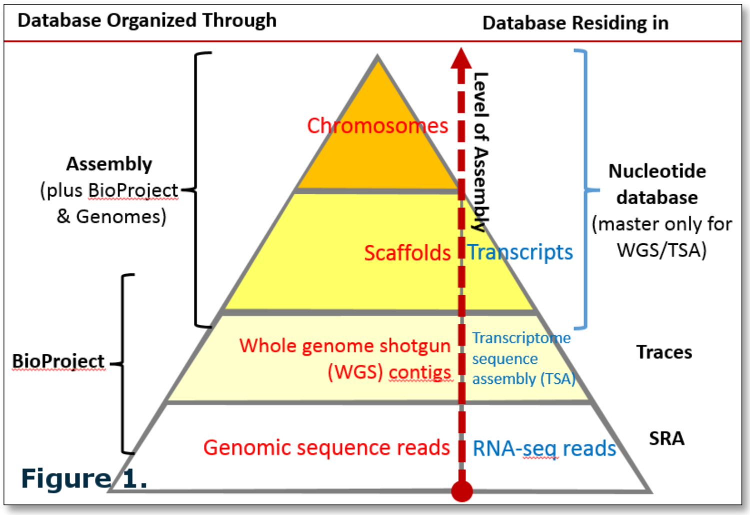
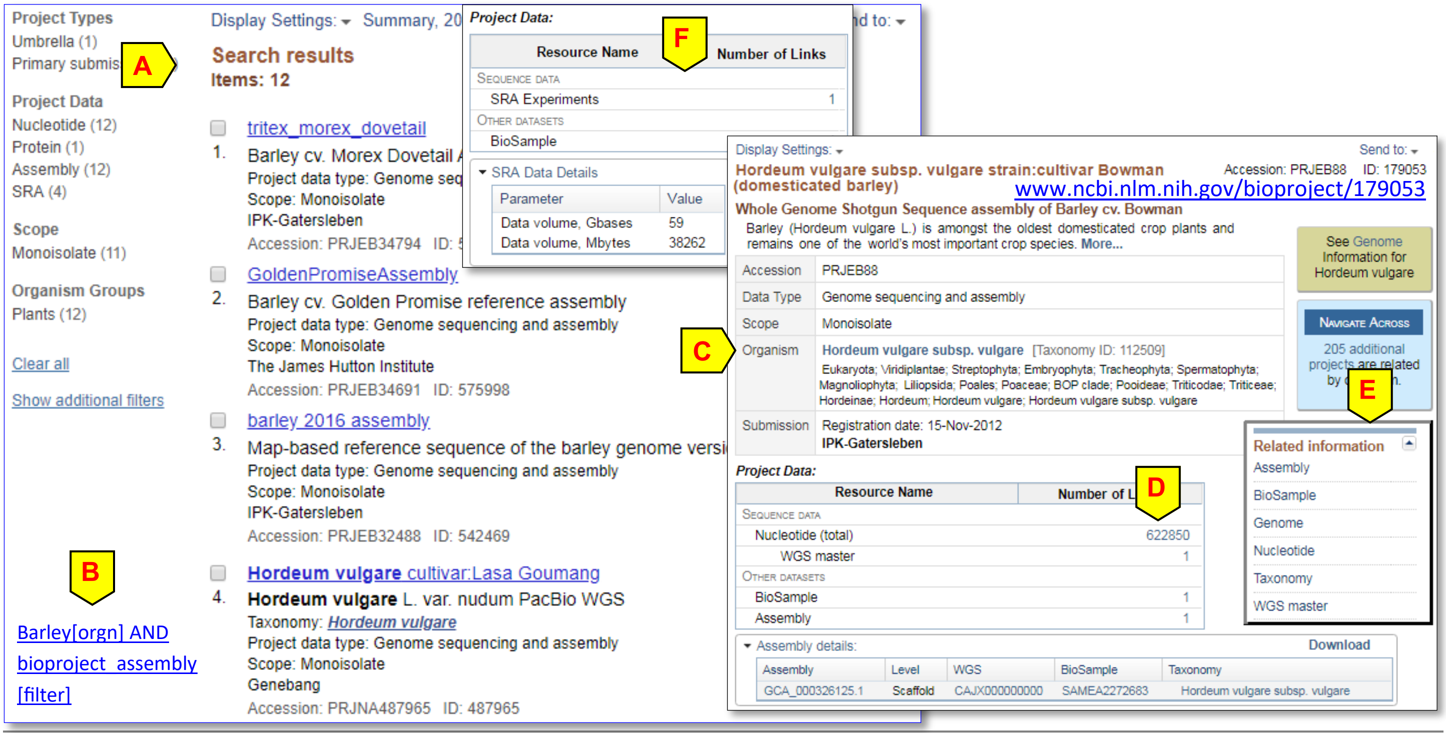

Database Organized Through
Database Residing in
Assembly (plus BioProject & Genomes)
BioProject
Chromosomes
Scaffolds
Transcripts
Whole genome shotgun (WGS) contigs
Transcriptome sequence assembly (TSA)
Genomic sequence reads
RNA-seq reads
Level of Assembly
Nucleotide database (master only for WGS/TSA)
Traces
SRA

Figure 1.

BioProject and Assembly Entries with Genomic Data

A BioProject database record provides a summary of a specific research project and lists all data available from the project. The result below (A) is from searching with "Barley[orgn] AND bioproject_assembly[filter]" (B). Click the title to open a record (C) for more details. The Project Data table lists available nucleotide sequences, with the number linked to actual records (wgs contigs in this case, D). The right hand column lists related records in other NCBI databases (insert, E). For project with only raw sequence reads, the link will point to Sequence Read Archive (SRA, F).



Project Types
Umbrella (1)
Primary submis

Project Data
Nucleotide (12)
Protein (1)
Assembly (12)
SRA (4)

Scope
Monoisolate (11)

Organism Groups
Plants (12)

Barley[orgn] AND bioproject_assembly [filter]

Search results
Items: 12

1. tritex_morex_dovetail
Barley cv. Morex Dovetail A
Project data type: Genome seq
Scope: Monoisolate
IPK-Gatersleben
Accession: PRJEB34794 ID: 5

2. GoldenPromiseAssembly
Barley cv. Golden Promise reference assembly
Project data type: Genome sequencing and assembly
Scope: Monoisolate
The James Hutton Institute
Accession: PRJEB34691 ID: 575998

3. barley 2016 assembly
Map-based reference sequence of the barley genome versi
Project data type: Genome sequencing and assembly
Scope: Monoisolate
IPK-Gatersleben
Accession: PRJEB32488 ID: 542469

4. Hordeum vulgare cultivar:Lasa Goumang
Hordeum vulgare L. var. nudum PacBio WGS
Taxonomy: *Hordeum vulgare*
Project data type: Genome sequencing and assembly
Scope: Monoisolate
Genebang
Accession: PRJNA487965 ID: 487965

Project Data:

Resource Name	Number of Links
Sequence data	
SRA Experiments	1
Other datasets	
BioSample	

SRA Data Details

Parameter	Value
Data volume, Gbases	59
Data volume, Mbytes	38262

Hordeum vulgare subsp. vulgare strain:cultivar Bowman (domesticated barley) Accession: PRJEB88 ID: 179053

www.ncbi.nlm.nih.gov/bioproject/179053

Whole Genome Shotgun Sequence assembly of Barley cv. Bowman

Barley (Hordeum vulgare L.) is amongst the oldest domesticated crop plants and remains one of the world's most important crop species. More...

Accession	PRJEB88
Data Type	Genome sequencing and assembly
Scope	Monoisolate
Organism	**Hordeum vulgare subsp. vulgare** [Taxonomy ID: 112509] Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BOP clade; Pooideae; Triticodae; Triticeae; Hordeinae; Hordeum; Hordeum vulgare; Hordeum vulgare subsp. vulgare
Submission	Registration date: 15-Nov-2012 **IPK-Gatersleben**

Project Data:

Resource Name	Number of L
Sequence data	
Nucleotide (total)	622850
WGS master	1
Other datasets	
BioSample	1
Assembly	1

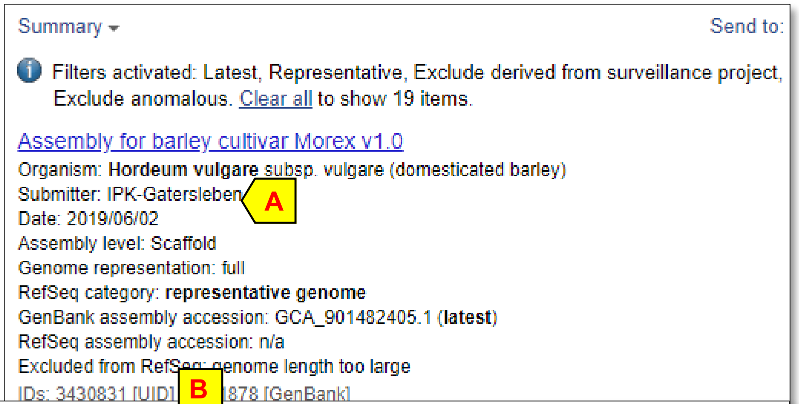
Assembly details:

Assembly	Level	WGS	BioSample	Taxonomy
GCA_000326125.1	Scaffold	CAJX00000000	SAMEA2272683	Hordeum vulgare subsp. vulgare

Related information
Assembly
BioSample
Genome
Nucleotide
Taxonomy
WGS master

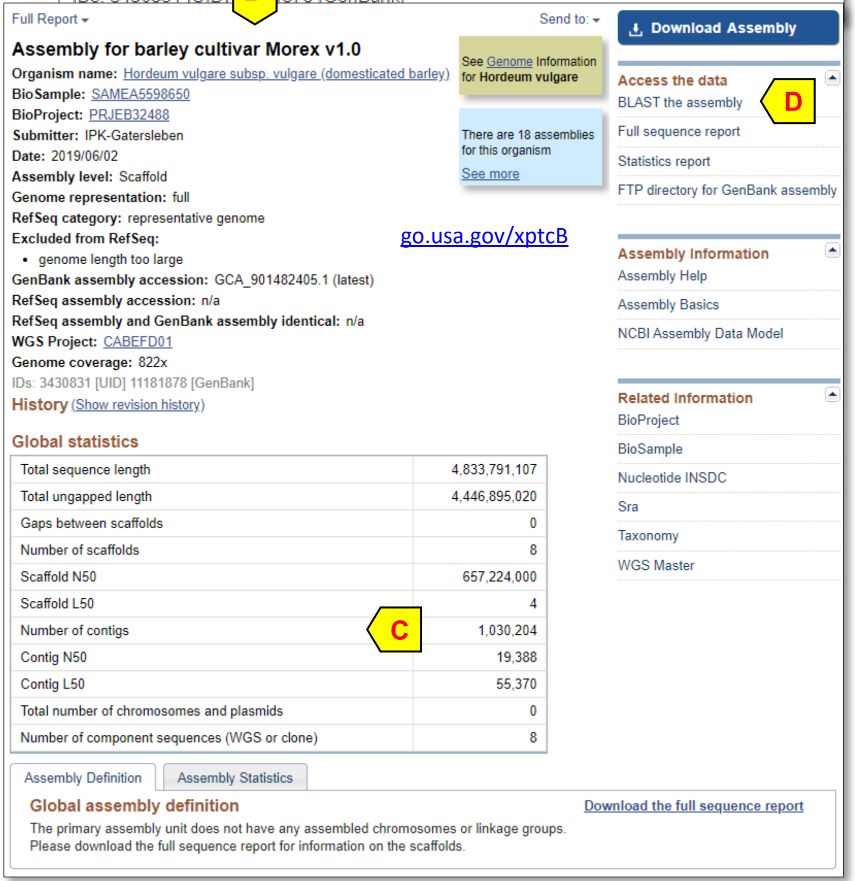
BioProject and Assembly Entries with Genomic Data (cont.)

An assembly database record provides summary information for a specific genomic assembly. Searching with “barley[orgn]” and filtering for the “Representatives” retrieves a single record (**A**). Its full report (**B**) contains a detailed description of the assembly, with chromosomal level details given in the “Global assembly definition” table (**C**) when available (absent for this specific assembly). The “BLAST the assembly” link in the right hand column (**D**) leads to a Assembly-specific BLAST search form.



Filters activated: Latest, Representative, Exclude derived from surveillance project, Exclude anomalous. Clear all to show 19 items.

Assembly for barley cultivar Morex v1.0
Organism: **Hordeum vulgare** subsp. vulgare (domesticated barley)
Submitter: IPK-Gatersleben
Date: 2019/06/02
Assembly level: Scaffold
Genome representation: full
RefSeq category: **representative genome**
GenBank assembly accession: GCA_901482405.1 (**latest**)
RefSeq assembly accession: n/a
Excluded from RefSeq: genome length too large
IDs: 3430831 [UID] 11181878 [GenBank]



Assembly for barley cultivar Morex v1.0
Organism name: Hordeum vulgare subsp. vulgare (domesticated barley)
BioSample: SAMEA5598650
BioProject: PRJEB32488
Submitter: IPK-Gatersleben
Date: 2019/06/02
Assembly level: Scaffold
Genome representation: full
RefSeq category: representative genome
Excluded from RefSeq:
- genome length too large

GenBank assembly accession: GCA_901482405.1 (latest)
RefSeq assembly accession: n/a
RefSeq assembly and GenBank assembly identical: n/a
WGS Project: CABEFD01
Genome coverage: 822x
IDs: 3430831 [UID] 11181878 [GenBank]
History (Show revision history)

See Genome Information for **Hordeum vulgare**

There are 18 assemblies for this organism
See more

go.usa.gov/xptcB

Access the data
BLAST the assembly
Full sequence report
Statistics report
FTP directory for GenBank assembly

Assembly Information
Assembly Help
Assembly Basics
NCBI Assembly Data Model

Related Information
BioProject
BioSample
Nucleotide INSDC
Sra
Taxonomy
WGS Master

Global statistics

Total sequence length	4,833,791,107
Total ungapped length	4,446,895,020
Gaps between scaffolds	0
Number of scaffolds	8
Scaffold N50	657,224,000
Scaffold L50	4
Number of contigs	1,030,204
Contig N50	19,388
Contig L50	55,370
Total number of chromosomes and plasmids	0
Number of component sequences (WGS or clone)	8

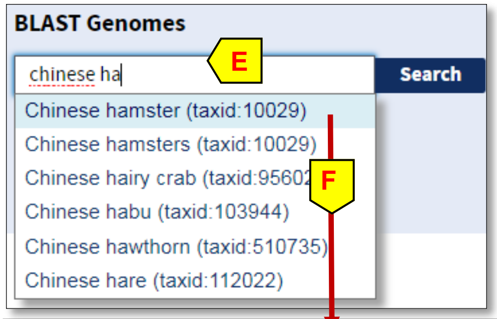
Global assembly definition
Download the full sequence report
The primary assembly unit does not have any assembled chromosomes or linkage groups.
Please download the full sequence report for information on the scaffolds.

Organism Search Box in the BLAST Homepage

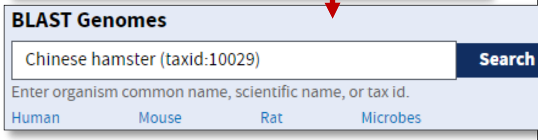
The search box in the BLAST Genome section of the BLAST homepage (blast.ncbi.nlm.nih.gov/) streamlines the process to allow quick access to the best genomic dataset for the input organism. It returns organism-specific nucleotide BLAST pages in the following decreasing level of assembly:

- fully annotated RefSeq chromosomal assembly (go.usa.gov/xptx4)
- Scaffold level of assembly, generally with annotation and some genomic context, but without chromosome-level placement
- WGS level of assembly, with or without annotation, without genomic context or chromosome-level placement
- NT database with the input organism as limit, if all the above fail

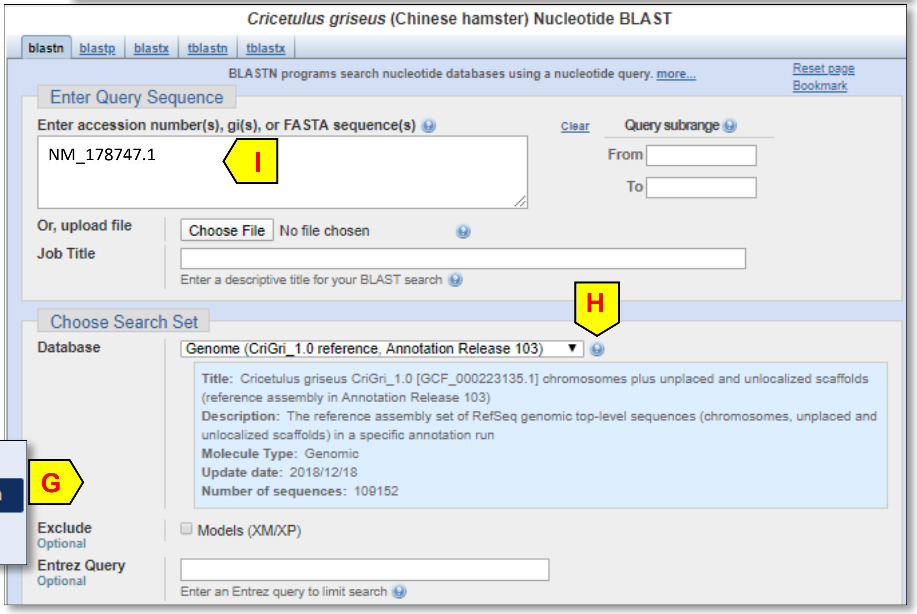
The example below locates the genome assembly for the Chinese hamster. It returns the organism-specific BLAST page with annotated RefSeq genome as the target database: type the organism name in the input box to see a suggested list (**E**), select the desired entry from the list (**F**), click “Search” to get to the search page (**G**), and click the “?” icon (**H**) to toggle on a detailed description of the selected database. The example uses the mouse mRNA of the vitamin C synthesis gene (NM_178747.1, **I**) to identify the hamster counterpart.



chinese ha
Chinese hamster (taxid:10029)
Chinese hamsters (taxid:10029)
Chinese hairy crab (taxid:95602)
Chinese habu (taxid:103944)
Chinese hawthorn (taxid:510735)
Chinese hare (taxid:112022)



Chinese hamster (taxid:10029)
Enter organism common name, scientific name, or tax id.



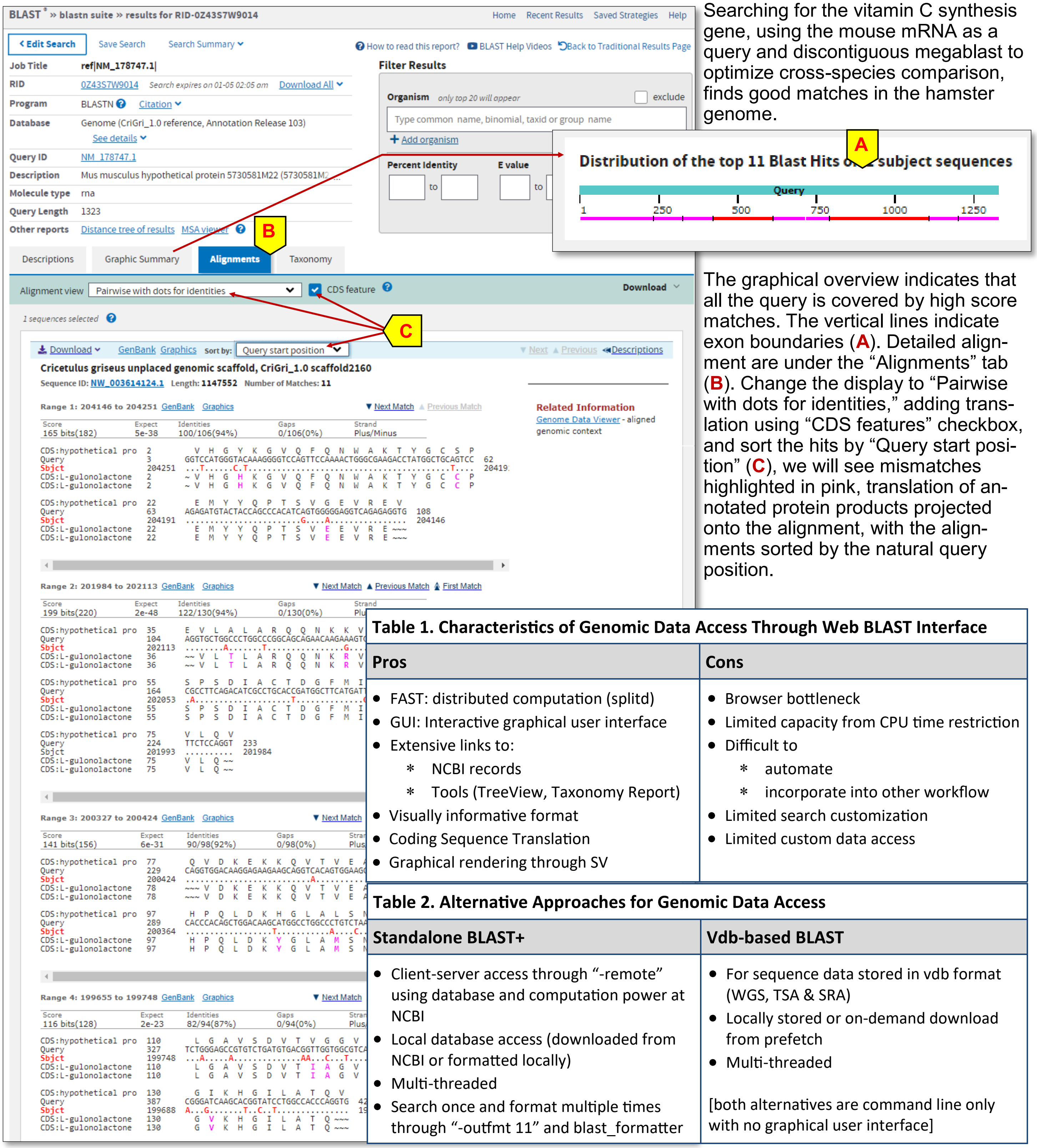
Cricetulus griseus (Chinese hamster) Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
NM_178747.1
Query subrange
From
To
Or, upload file: No file chosen
Job Title
Enter a descriptive title for your BLAST search

Choose Search Set
Database: Genome (CriGri_1.0 reference, Annotation Release 103)
Title: Cricetulus griseus CriGri_1.0 [GCF_000223135.1] chromosomes plus unplaced and unlocalized scaffolds (reference assembly in Annotation Release 103)
Description: The reference assembly set of RefSeq genomic top-level sequences (chromosomes, unplaced and unlocalized scaffolds) in a specific annotation run
Molecule Type: Genomic
Update date: 2018/12/18
Number of sequences: 109152
Exclude (Optional): □ Models (XM/XP)
Entrez Query (Optional)
Enter an Entrez query to limit search

An Example Web BLAST Search Result



BLAST ® » blastn suite » results for RID-0Z43S7W9014

Home Recent Results Saved Strategies Help

< Edit Search Save Search Search Summary

How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title	ref|NM_178747.1|
RID	0Z43S7W9014 *Search expires on 01-05 02:05 am* Download All
Program	BLASTN Citation
Database	Genome (CriGri_1.0 reference, Annotation Release 103) See details
Query ID	NM_178747.1
Description	Mus musculus hypothetical protein 5730581M22 (5730581M2 ...
Molecule type	rna
Query Length	1323
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to E value to

Distribution of the top 11 Blast Hits on 12 subject sequences

Descriptions Graphic Summary Alignments Taxonomy

Alignment view Pairwise with dots for identities ☑ CDS feature Download

1 sequences selected

Download GenBank Graphics Sort by: Query start position ▼ Next ▲ Previous ◀Descriptions

Cricetulus griseus unplaced genomic scaffold, CriGri_1.0 scaffold2160

Sequence ID: NW_003614124.1 Length: 1147552 Number of Matches: 11

Related Information

Genome Data Viewer - aligned genomic context

Range 1: 204146 to 204251 GenBank Graphics ▼ Next Match ▲ Previous Match

```
Score           Expect   Identities      Gaps       Strand
165 bits(182)   5e-38    100/106(94%)    0/106(0%)  Plus/Minus

CDS:hypothetical pro  2         V  H  G  Y  K  G  V  Q  F  Q  N  W  A  K  T  Y  G  C  S  P
Query                 3        GGTCCATGGGTACAAAGGGGTCCAGTTCCAAAACTGGGCGAAGACCTATGGCTGCAGTCC  62
Sbjct                 204251   ...T......C.T........................................T....  20419
CDS:L-gulonolactone   2        ~ V  H  G  H  K  G  V  Q  F  Q  N  W  A  K  T  Y  G  C  C  P
CDS:L-gulonolactone   2        ~ V  H  G  H  K  G  V  Q  F  Q  N  W  A  K  T  Y  G  C  C  P

CDS:hypothetical pro  22        E  M  Y  Y  Q  P  T  S  V  G  E  V  R  E  V
Query                 63       AGAGATGTACTACCAGCCCACATCAGTGGGGAGGTCAGAGAGGTG  108
Sbjct                 204191   ........................G....A...............  204146
CDS:L-gulonolactone   22        E  M  Y  Y  Q  P  T  S  V  E  E  V  R  E ~~~
CDS:L-gulonolactone   22        E  M  Y  Y  Q  P  T  S  V  E  E  V  R  E ~~~
```

Range 2: 201984 to 202113 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

```
Score           Expect   Identities      Gaps       Strand
199 bits(220)   2e-48    122/130(94%)    0/130(0%)  Plu

CDS:hypothetical pro  35       E  V  L  A  L  A  R  Q  Q  N  K  K  V
Query                 104      AGGTGCTGGCCCTGGCCCGGCAGCAGAACAAGAAAGT
Sbjct                 202113   ........A.......T................G...
CDS:L-gulonolactone   36       ~~ V  L  T  L  A  R  Q  Q  N  K  R  V
CDS:L-gulonolactone   36       ~~ V  L  T  L  A  R  Q  Q  N  K  R  V

CDS:hypothetical pro  55       S  P  S  D  I  A  C  T  D  G  F  M  I
Query                 164      CGCCTTCAGACATCGCCTGCACCGATGGCTTCATGATT
Sbjct                 202053   .A.....................T..............
CDS:L-gulonolactone   55       S  P  S  D  I  A  C  T  D  G  F  M  I
CDS:L-gulonolactone   55       S  P  S  D  I  A  C  T  D  G  F  M  I

CDS:hypothetical pro  75       V  L  Q  V
Query                 224      TTCTCCAGGT  233
Sbjct                 201993   ..........  201984
CDS:L-gulonolactone   75       V  L  Q ~~
CDS:L-gulonolactone   75       V  L  Q ~~
```

Range 3: 200327 to 200424 GenBank Graphics ▼ Next Match

```
Score           Expect   Identities      Gaps       Stran
141 bits(156)   6e-31    90/98(92%)      0/98(0%)   Plus

CDS:hypothetical pro  77        Q  V  D  K  E  K  K  Q  V  T  V  E  A
Query                 229      CAGGTGGACAAGGAGAAGAAGCAGGTCACAGTGGAAG
Sbjct                 200424   .......................A.............
CDS:L-gulonolactone   78       ~~~ V  D  K  E  K  K  Q  V  T  V  E  A
CDS:L-gulonolactone   78       ~~~ V  D  K  E  K  K  Q  V  T  V  E  A

CDS:hypothetical pro  97        H  P  Q  L  D  K  H  G  L  A  L  S  N
Query                 289      CACCCACAGCTGGACAAGCATGGCCTGGCCCTGTCTAA
Sbjct                 200364   ..................T..........A....C..
CDS:L-gulonolactone   97        H  P  Q  L  D  K  Y  G  L  A  M  S  N
CDS:L-gulonolactone   97        H  P  Q  L  D  K  Y  G  L  A  M  S  N
```

Range 4: 199655 to 199748 GenBank Graphics ▼ Next Match

```
Score           Expect   Identities      Gaps       Stran
116 bits(128)   2e-23    82/94(87%)      0/94(0%)   Plus

CDS:hypothetical pro  110       L  G  A  V  S  D  V  T  V  G  G  V
Query                 327      TCTGGGAGCCGTGTCTGATGTGACGGTTGGTGGCGTCA
Sbjct                 199748   ...A.....A.............AA...C...T....
CDS:L-gulonolactone   110       L  G  A  V  S  D  V  T  I  A  G  V
CDS:L-gulonolactone   110       L  G  A  V  S  D  V  T  I  A  G  V

CDS:hypothetical pro  130       G  I  K  H  G  I  L  A  T  Q  V
Query                 387      CGGGATCAAGCACGGTATCCTGGCCACCCAGGTG  42
Sbjct                 199688   A...G.......T..C..T...............  19
CDS:L-gulonolactone   130       G  V  K  H  G  I  L  A  T  Q ~~~
CDS:L-gulonolactone   130       G  V  K  H  G  I  L  A  T  Q ~~~
```

Searching for the vitamin C synthesis gene, using the mouse mRNA as a query and discontiguous megablast to optimize cross-species comparison, finds good matches in the hamster genome.

The graphical overview indicates that all the query is covered by high score matches. The vertical lines indicate exon boundaries (**A**). Detailed alignment are under the “Alignments” tab (**B**). Change the display to “Pairwise with dots for identities,” adding translation using “CDS features” checkbox, and sort the hits by “Query start position” (**C**), we will see mismatches highlighted in pink, translation of annotated protein products projected onto the alignment, with the alignments sorted by the natural query position.

Table 1. Characteristics of Genomic Data Access Through Web BLAST Interface

Pros	Cons
• FAST: distributed computation (splitd) • GUI: Interactive graphical user interface • Extensive links to: * NCBI records * Tools (TreeView, Taxonomy Report) • Visually informative format • Coding Sequence Translation • Graphical rendering through SV	• Browser bottleneck • Limited capacity from CPU time restriction • Difficult to * automate * incorporate into other workflow • Limited search customization • Limited custom data access

Table 2. Alternative Approaches for Genomic Data Access

Standalone BLAST+	Vdb-based BLAST
• Client-server access through “-remote” using database and computation power at NCBI • Local database access (downloaded from NCBI or formatted locally) • Multi-threaded • Search once and format multiple times through “-outfmt 11” and blast_formatter	• For sequence data stored in vdb format (WGS, TSA & SRA) • Locally stored or on-demand download from prefetch • Multi-threaded [both alternatives are command line only with no graphical user interface]

Web Access: Some Pros & Cons

BLAST access of genomic sequence data through the web interface has advantages and limitations (Table 1). For genomic BLAST searches requiring high throughput, customization, or workflow integration, alternative approaches may work better. These alternative approaches include the standalone BLAST+ package, the vdb-based BLAST programs from the NCBI sratoolkit, as well as the cloud implementation. Table 2 summarizes some of the characteristics for BLAST+ and vdb-based BLAST tools.

Locate the Genomic Dataset Using blastdbinfo

The blastdbinfo database provides information about BLAST databases available from NCBI. We can use the einfo tool from the Entrez Program Utilities API to get the basic information about this database (go.usa.gov/xptgr). To locate a database, we can search with text terms through esearch, then refer to the uids (or the **WebEnv** and **QueryKey** value when using **&usehistory=y**) to get details through esummary. For BLAST+ programs, we can use the value from the **<Path>** field as an argument for the **-db** switch in combination with **-remote** to search against that dataset. The example (right, **A**) searches for the wine grape genomic assembly and parses the <Path> values using the XML parser xtract from EDirect (go.usa.gov/xptgb). The value serves as an argument in a blastn search (**B**).

```
eDirect Commands
esearch -db blastdbinfo -query "wine grape[database organism taxid] AND genomic[blast database type] AND
refseq[blast database source]" | esummary | xtract -pattern DocumentSummary -element Name,Path

OUTPUT
RefSeq Genomic  GP/29760.12992/RefSeq_Genomic                                                      A
… …
GCF_000003745.3 genomic/29760/GCF_000003745.3 <<<<<<<<<<Path value we need
                                                          B
blast+ commands
blastn -db genomic/29760/GCF_000003745.3 -remote -query my_query.fa <other switches>
```

As with other Entrez databases, we can use field-limited terms to do more precise searches as shown below.

- genbank[Blast Database Source], gnomon, pdb, refseq, sra, swissprot, trace
- wgs[Blast Sequence Strategy]
- cdna[Blast Sequence Type], genomic, otherdna, protein
- gca_000003225_1[Genome Collection Assembly Name], gcf_000181295_1
- alistipes_inops[Database Organism Taxid], similar to [organism] in other Entrez databases
- 104937[NCBI Genome Project ID]
- aaaa[NCBI WGS Project ID]
- nucleotide[database title]

For sequence data stored in vdb format (WGS, TSA, and SRA), we need to use **blastn_vdb** and **tblastn_vdb** programS from the SRA toolkit to perform searches. For historical reasons, blastdbinfo's <Path> value for WGS and TSA entries contains a "*WGS_VDB://*" prefix that must be removed before using the value as argument for the **-db** switch. The example above (**C**) searches three databases (space-separated within quotes) in vdb formats. For WGS datasets with many volumes, NCBI provides a tool, ***taxid2wgs.pl***, to facilitate the collection of WGS datasets available for a specific organism using its taxonomy id (taxid). The example collects available WGS datasets for *Streptomyces scabiei* using its taxid, 1930 (**D**). More information on this tool is available at ftp.ncbi.nlm.nih.gov/blast/WGS_TOOLS/

```
                                                    C
vdb blastn command
blastn_vdb -db "SRR011188 GACC01 AJKK01" -query q -outfmt '6 sseqid length evalue bitscore'
Tabular output
gi|425936258|gb|AJKK01254946.1| 180    9.56e-89     333
gi|425936258|gb|AJKK01254946.1| 125    3.59e-58     231
… …
gnl|SRA|SRR011188.63012.2     109    2.81e-49     202
gnl|SRA|SRR011188.21834.2     105    4.71e-47     195
gnl|SRA|SRR011188.27702.2     104    1.69e-46     193
… …                                                                                D
taxid2wgs.pl example:
perl taxid2wgs.pl -alias_file "Streptomyces_scabiei_WGS" -title "Streptomyces scabiei WGS"
1930

Content of the vdb database alias file
#
# Alias file created by taxid2wgs.pl/1.0 on Sat Dec 26 11:12:03 2015
TITLE Streptomyces scabiei WGS
VDBLIST JPPX01 JPPW01 LBNJ01
```

Download Databases

We can install BLAST databases locally by downloading them from the NCBI ftp site. Table 3 summarizes the available databases, their formats, and FTP subdirectories.

Table 3. Alternative Approaches for Genomic Data Access

Databases	Details
Standard set (preformatted)	ftp.ncbi.nlm.nih.gov/blast/db/ Download all volumes, e.g., nt.##.tar.gz, Extract and go
RefSeq genome assemblies (FASTA)	ftp.ncbi.nlm.nih.gov/genomes/refseq/ Refer to the readme
WGS+TSA * (vdb)	ftp.ncbi.nlm.nih.gov/sra/wgs/ Locate project initials using Entrez Nucleotide query: wheat [orgn] AND (wgs_master[prop] OR tsa_master[prop])
WGS + TSA (FASTA)	www.ncbi.nlm.nih.gov/Traces/wgs/ Browse and navigate to individual project to download
SRA (vdb) *	ftp.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/ Search in BioProject & follow links to SRA to locate the SRR#

** Using prefetch from the sratoolkit to download datasets in vdb format is strongly recommended!*

Other Help Documents

Standalone BLAST+
go.usa.gov/xpt4a

Local vdb_blast:
go.usa.gov/xpt4c

BLAST in the Cloud
github.com/ncbi/blast_plus_docs